Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 333–339

Procedia Computer Science

www.elsevier.com/locate/procedia

21st International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Linguistic Data Classification with Combined Comparison Measures

Kalle Saastamoinen

National Defence University, P.O. Box 7, Helsinki FI-00861 , Finland

Abstract

Medical data is often imprecise, due to many reasons that can be technical or human originated. In this article, we will present a classification example where data in hand is given imprecisely. Data set presents a choice situation where medical doctor has to be able to make a decision where patient is to be sent after the surgery. Data is given linguistically, which might give the idea to use some kind of fuzzy numbers in order to decode linguistic variable into the classifiable form. In fact, this approach makes the data more imprecise and therefore harder to classify. On another hand finding of parameter values by the use of commonly used differential evolution (DE) is very time consuming. In this article, we use simple, yet effective method for decoding of linguistic data. After this we use randomly selected weights and t-norm based combined comparison measures with similarity classifier to classify data given to the correct classes. Results are compared to the existing results and method presented in this paper provides best total rate of true positive classification result of 88.89% using combination of Yager t-norm and t-conorm, whereas second highest reported best total rate of true positive classification result was 77.27% using similarity measure called Shweizer & Sklar - Łukasiewicz and Differential Evolution (DE).

Peer-review under responsibility of KES International

Keywords: Linguistic Data; Post-operative; Classification

1. Introduction

In our previous article[1] we found very good results using simple method for decoding of linguistic data and Shweizer & Sklar - Łukasiewicz similarity based comparison measure with Post-operative data[2]. In that article, we assumed that testing with different comparison measures and step-sizes would likely make results better. This article shows that assumptions were correct ones.

Comparison is very important task in all areas where decisions has to be done. The fields of problem solving, categorization, data mining, classification, memory retrieval, inductive reasoning, and generally cognitive processes

* Corresponding author. Tel.: +358407606489.
E-mail address: kalle.saastamoinen@mil.fi

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.005

require that we understand how to assess the sameness. Measures used for comparison can in general have many different forms depending the purpose of their utilization[3].

Best result using randomized weights and measure created from combination of Yager t-norm and t-conorm was 88.89%, while previous highest was 77.27% in[1] and before this it was 71.11% in[4]. On another hand, best average results were always achieved using DE instead of randomized weights. Best average result now was achieved by the use of combined Yager measure with DE that was 81.82% while previous best average results had been 65.23% in[1] and before this 62.67% in[4].

In this article, it is used a comparison measure that originates from many valued logical structures called t-norms and t-conorms. Fuzzy logic has traditionally been used to transform linguistic data into the numerical values that are easy to handle with some computing system like for example Matlab. This transformation is usually carried out by using membership functions, which the most common ones are trapezoidal, triangular and Gaussian, where their names refers to their shapes. Idea of using membership functions is to represent data in a set where data points takes some membership degrees which again represents in what degree these points belongs in the sets which are characteristic for the problem in hand. Commonly if one uses membership functions one also need to use some defuzzification method in order to produce some quantifiable results. Interesting approach to classify Post-operative data[2] using this kind of method was presented in[4]. Here it is presented simple approach, which does not need membership functions nor complicated defuzzification to classify Post-operative data set and it proves to give better results than any method has been able to shown before. The core of the used method is similarity measure that is based on Yager class of t-norms and t-conorms[5] and likely the simplest meaningful transformation of Post-operative linguistic data into the numeric values that corresponds intuitively reality.

Motivation for this article is to present how previous results presented in article[1] can get better by the use of other comparison measures besides Shweizer & Sklar - Łukasiewicz[1], when dealing with Post-operative type of linguistic data.

Article is organized as follows. In the first section, Post-operative data and of how this linguistics data has been handled is presented. The second section gives a little theory behind comparison measures used. The third section presents classification schemata. The fourth section presents results achieved and these results are compared vs. to the previous results achieved. In the fifth section conclusions are done and some future directions are given.

2. Post-operative data

Task of this data is to determine where patients in a postoperative recovery area should be sent to next. The attributes correspond roughly to the body temperature measurements. The number of Instances is 90. The number of attributes is 9 including the decision (class attribute). Attribute 8 has 3 missing values[2]. So each patient is defined by 8 attribute values, from which one should be able to make a decision which of 3 recovery areas (marked as 3 different classes) patient should be sent.

What makes this data challenging is that most of the attribute values are given by linguistic labels. Three attributes that gets values low, mid and high are patient's internal temperature (named as L-CORE), surface temperature (named as L-SURF) and last measurement of blood pressure (named as L-BP). Patient's oxygen saturation (named as L-O2) gets values excellent, good, fair and poor. Three attributes that gets values stable, mod-stable and unstable are stability of patient's surface temperature (named as SURF-STBL), stability of patient's core temperature (named as CORE-STBL) and stability of patient's blood pressure (named as BP-STBL). Attribute that describes patient's perceived comfort at discharge (named as COMFORT) is measured as an integer between 0 and 20. Last attribute describes discharge decision (named as decision ADM-DECS) it can get 3 values that are I (patient sent to Intensive Care Unit), S (patient prepared to go home) and A (patient sent to general hospital floor).

Comfort attribute has 3 missing values, which increases uncertainty in Post-operative. Furthermore there is no information what values experts used exactly to set up linguistic labels for the attribute values. It is only given some limits like for L-CORE high (> 37), mid (≥ 36 and ≤ 37), low (< 36). However no data is given what exactly patient's internal temperature was, when it is labeled as example to be Mid. For this reason these limits are useless. In fact it only creates more uncertainty if one starts to guess these values as it is done in case one starts to use some membership functions. Below it is shown an example of Post-operative data concerning first patient: